International Journal of Computer Science, Engineering and Information Technology (IJCSEIT),
Vol.12, No.2/3/4, August 2022

DATA WAREHOUSING AS KNOWLEDGE POOL : A VITAL COMPONENT OF BUSINESS INTELLIGENCE

Abobakr Aljuwaiber

Department of Administrative and Financial Sciences and Technologies,
Applied College, Taibah University, Saudi Arabia

ABSTRACT

Increasing amounts of information and diverse formats have forced organizations to create large data repositories in response to the information explosion in the 21st century. As a result, the model of a data warehouse has been introduced to define a large data repository. The purpose of this article is to describe the principles of data warehousing in business and how it can enhance the generation of new knowledge throughout the organization. Definitions of data warehousing are considered and include methods of its use, namely query and data simulation. The steps required prior to transforming the raw data into a data warehouse and project goals for data warehouses and metadata are also discussed. The objective of this document is to provide a clear and simple description of data warehousing terms and concepts, especially for busy managers and laymen. They may only need basic and direct information about data warehouses to gain a complete understanding of the principles of the data warehouse.

KEYWORDS

Data warehouse, Business, Metadata, Knowledge pool.

1. INTRODUCTION

In the business realm, obtaining the right information is crucial to making the right business decisions. Since the late mid-1990s, data warehouses have become one of the crucial developments in the field of information systems. Data warehousing is an increasingly popular and powerful concept for applying information technology to solve business problems. In fact, a data warehouse improves the productivity of business decisions by consolidating, transforming, and integrating operational data, providing a consistent view of an organization. It can be used as a knowledge pool for organizations where information and data can be retrieved and accessed [13]. Therefore, the data warehouse is an important knowledge source.

The data warehouse provides Online Analytical Processing (OLAP) tools. OLAP is described by Khan [10, p. 108] as "a business intelligence tool that addresses the need to perform multidimensional analysis". It is one of the functions that aid knowledge workers (executives, managers, and analysts) in analysing data of business transactions stored in the dimensional data warehouse, thus facilitating their tactical and strategic business decisions [18, 19]. Furthermore, OLAP is considered a vital part of business intelligence that deals with a huge amount of data [11]. Hence, the data warehouse is the core of the analytics process and is the instrument that assures that the information extracted, transformed and loaded meets the quality standards that are necessary to have a quality output in the generation of organizational knowledge. This concept can assist knowledge workers by providing them with an intelligent analysis platform to use as a decision support system [15, 16].

As new information and communication technologies emerged, companies needed to deal with big data analytics. Hence, data warehouses and database technologies were utilized to address this issue. Essentially, data warehousing assists the construction of a huge repository of integrated data that is enhanced for analytic purposes [8]. A data warehouse has one of the most important components: the metadata stored in the warehouse. Accordingly, the big challenge is how to extract meaningful value from the data set [6].

The main research question of this article is "What are the most important principles of data warehousing in business that can enhance the generation of new knowledge throughout the organization?". Related to this question, the objective of this paper is to provide readers with the most important aspects of data storage and data usage. The paper considers various methods of transforming and integrating data warehouses and their real-world applications. A recent study by Križanić & Rabuzin [11, p. 39] underlined the importance of applying various business process analyses to the domains of data warehouse applications in such sectors as health, higher education, and business. According to Chandra & Gupta [3, p. 6] stated that the applications of data warehousing are growing importance rapidly in the technology and business fields. Furthermore, size, speed, and parallel and distributed data warehouse are the thrilling field of research. Arising from curiosity about that topic, the main contribution of this paper is to give an overview of the utilization of raw data in the context of data warehouses for possible contribution and improvement to business within the domains of data warehousing and mining. Additionally, it proposes novel optimization processes that assist businesses by providing appropriate steps for transforming raw data into a data warehouse. The article is organized as follows: Section 2 begins with some definitions and uses of data warehouses. Section 3 illustrates steps for converting raw data into a data warehouse. Section 4 discusses the goals of the project of data warehousing. Finally, one of the most important components of a data warehouse, the so-called metadata, is described. The article concludes with some ideas for future research.

2. DEFINITION OF DATA WAREHOUSE

Data warehousing is a concept and not a product that can be bought off the shelf. Data warehouses are expensive to build and difficult to protect [19]. There are various definitions of a data warehouse; for instance, Rainardi [18, p. 1] defined a data warehouse as "a system that retrieves and consolidates data periodically from the source system into dimensional or normalized data store". Thus, it can be viewed as a database that collects data from a variety of sources within an organization, including the company's processes, products, and customers [17]. A data warehouse collects discrete data from different sources, structured and optimised for query access using an OLAP (Online Analytical Processing) query tool. Thus, OLAP techniques are commonly applied by data warehousing and process mining [11]. Chaudhuri and Dayal [4, p. 65] defined data warehousing as "a collection of decision support technologies, aimed at enabling the knowledge worker (executive, manager, analyst) to make better and faster decisions".

2.1. Why Data Warehousing?

The main idea of data warehousing comes from companies that have developed and implemented multiple information systems. Data warehouses are designed to store, extract, and transfer data for strategic analysis [7]. These processes allow the data warehouse to provide decision-makers with useful information that helps improve their business processes [14]. Thus, ad hoc queries and quick response time are considered important advantages of data warehouses [11]. However, security issues can be a source of concern for a data warehouse and should be considered in the field of information systems since a data warehouse is an open area [9].

2.2. Using Data Warehouses

There are two basic ways to access information in a data warehouse, which are queries and data mining.

1. Queries: A query can be defined as "the process of getting data from a data store, which satisfies certain criteria" [18, p. 11]. A data warehouse can be examined with query language such as Structured Query Language (SQL) or by using Online Analytical Processing (OLAP).
2. Data mining: Some research refers to data mining as knowledge discovery in database (KDD). Data mining is an experimental, exploratory, and iterative process that consists of a number of stages [21]. It helps end-users extract useful business information from a large database. Thus, data mining uses developed statistical processing or artificial intelligence programs to analyse data and find patterns [7].

3. STEPS FOR TRANSFORMING RAW DATA INTO A DATA WAREHOUSE

A data warehouse converts raw data into a useful analytical tool for business decision-making. Some companies import raw data into Online Transaction Processing (OLTP) systems, which follow day-to-day operations: for example, sales, purchases, and inventory change [16, 18]. This step is followed by integrated processes of transforming raw data into a data warehouse where various formats can be rigorously unified by the extract-transform-load (ETL) procedure into the database via extraction, consolidation, filtering, transformation, cleansing, conversion, and aggregation [5, 10, 12]. Figure 1 illustrates the process of getting data from original sources into the data warehouse:

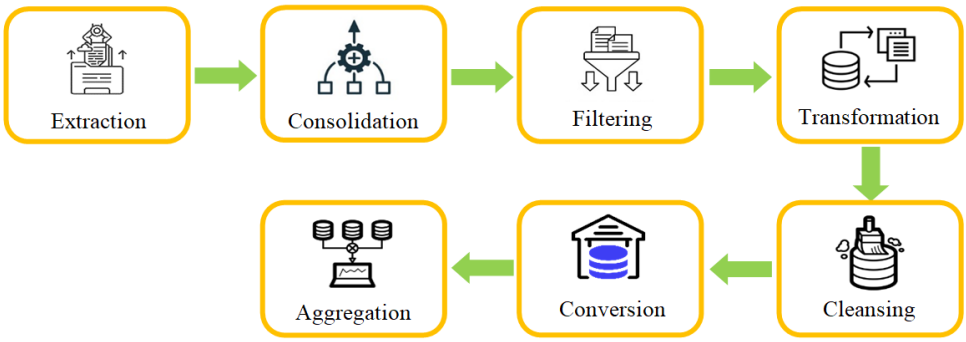

Figure 1. Processes of transforming raw data into a data warehouse

1. Extraction: This step usually deals with various data sources. In this step, the one who in charge may take data out of its original source in the database, then transform it into a data warehouse infrastructure.
2. Consolidation: This means the process of integrating data from multiple sources into one database.
3. Filtering: In this process, we take the useful data and discard the rest.
4. Transformation: Involves subjecting the data to a number of operations before importing it and deriving new values to be loaded to the target system, or for validating the data from the source system.
5. Cleansing: This step is important for ensuring high data quality. which improves the accuracy of the data in the warehouse. However, a large volume of data in the warehouse may have errors or anomalies due to the multiple sources involved [4], such as inconsistent

field lengths, inconsistent descriptions, inconsistent value assignments, missing entries, and violation of integrity constraints. Hence, it is important that data in the warehouse be correct.

6. Conversion: This step involves planning the raw data onto a new data field inside the warehouse, including typing and translating the data into the format used by a warehouse.
7. Aggregation: In this last step, many times the value of a data warehouse is in the abbreviated and extracted data it contains; the aggregation step classifies and joins data into useful metrics for analysis.

4. OBJECTIVES OF A DATA WAREHOUSE PROJECT

The most important way to achieve the objectives of a project is to have team members who come from different backgrounds and have been trained in different disciplines. The team should include members with different academic backgrounds or majors in different fields. For example, the team might comprise business specialists, technical systems specialists, data modellers, data analysts, and database administrators who are working together to achieve the same goal [17].
In the meantime, team members must strive to pay attention to and respond to customer needs, which is very important in global markets. When they know a customer's desire this will help them to deal with every customer with the individual care the customer demands [8, 17].

There are four essential objectives for almost any data-warehousing project:

- Develop a coordinated and collaborative approach toward data collection.
- Institute a consistent development approach and methodology for data warehousing.
- Capitalize on existing data warehouse experience and knowledge.
- Develop common terminology and concepts.

Beyond these requirements, a more fundamental question may be whether the team understands what information architecture is and whether the data warehouse project has a unified structure in place [17].

5. METADATA

Metadata can be defined as one of the most important aspects of data warehousing. It is "data about data" that describes and is stored in the warehouse and has several uses [1]. Metadata is vital for building, maintaining, managing, and using a data warehouse. Moreover, it provides useful information for locating data stored in the data warehouse [2]. Metadata delivers some useful components, including the following:

1. The location and description of the warehouse system and data components.
2. Names, definition, structure, and content of data warehouse and end observation.
3. Verification of original data sources.
4. For data to remain in the warehouse, consolidation and transformation rules are used, and these include mapping methods from the operational database.
5. Integration and transformation rules are used to deliver data to end-user analytical tools.
6. Data warehouse operational information which includes a history of warehouse updates, versions, and ownership authorizations [2, 10].

5.1. Metadata Trends

The trend toward integrating external data within data warehouses is clearly one of the best ways to model data warehouses. Thus, metadata trends have processes of integration of external and internal data into a warehouse that are illustrated in the following: (1) different data formats, (2) missing or invalid data, (3) levels having different aggregation, (4) indicative inconsistency, and (5) unknown data quality and timeliness. All these processes put an additional load on the collection and management of the general metadata [1, 10].

6. CONCLUSIONS

Increasing amounts of information and diverse formats have forced organizations to create large data repositories in response to the information explosion in the 21st century. As a result, the concept of a data warehouse has been introduced to define a large data repository. This article began with a discussion of the definition of data warehousing as the collection of data from multiple sources and the provision of that data to end users in a consolidated, consistent manner. It is essential for companies to improve multiple information systems in the data warehousing concept. The process of converting raw data into a useful data warehouse, which includes extraction, consolidation, filtering, transformation, cleansing, conversion, and aggregation, has also been illustrated. The success of a data warehouse project depends on the activities of its team members and their success in achieving the necessary goals and plans [17]. Metadata is one of the most important components of a data warehouse. Knowledge of metadata trends is necessary for a business to reduce costs and rapidly increase its competitiveness. Therefore, it is important for businesses to consider the rapid evolution of information technology and communications, and how to employ big data technology to solve the problems and changes affecting the modern data warehouse. Future research could be oriented toward examining the steps for transforming raw data into a data warehouse, as described in figure 1, in different sectors such as health and higher education, taking advantageous and disadvantageous aspects of the application into consideration. Future studies could also be conducted to expand upon this work by evaluating possible measures for improvement, such as the prevention of data leakage.

REFERENCES

[1] Berson, A., & Smith, S. J. (1997). Data warehousing, data mining, and OLAP. McGraw-Hill, Inc..
[2] Bhansali, N. (2009). Strategic data warehousing: achieving alignment with business. Auerbach Publications.
[3] Chandra, P., & Gupta, M. K. (2018). Comprehensive survey on data warehousing research. International Journal of Information Technology, Vol. 10, No. 2, pp 217-224.
[4] Chaudhuri, S., & Dayal, U. (1997). An overview of data warehousing and OLAP technology. ACM Sigmod record, Vol. 26, No.1, pp 65-74.
[5] Coronel, C., & Morris, S. (2016). Database systems: design, implementation, & management. Cengage Learning.
[6] Dabbèchi, H., Nabli, A., & Bouzguenda, L. (2016, September). Towards cloud-based data warehouse as a service for big data analytics. In International Conference on Computational Collective Intelligence (pp. 180-189). Springer, Cham.
[7] David, J. S., &Steinbart, P. J. (1999). Drowning in data. Strategic finance, Vol. 81, No. 6, pp 30-34.
[8] Garani, G., Chernov, A., Savvas, I., & Butakova, M. (2019, June). A data warehouse approach for business intelligence. In 2019 IEEE 28th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE) (pp. 70-75). IEEE.
[9] Gosain, A., & Arora, A. (2015). Security issues in data warehouse: a systematic review. Procedia Computer Science, Vol. 48, pp 149-157.
[10] Khan, A. (2012). Business Intelligence and data warehousing simplified: 500 questions, answers, and tips. Mercury Learning and Information.

[11] Križanić, S., & Rabuzin, K. (2020). Process mining and data warehousing-a literature review. In Central European Conference on Information and Intelligent Systems (pp. 33-39). Faculty of Organization and Informatics Varazdin.
[12] Lin, H. C. (2012). Real-time clinical decision support system. In Medical Informatics, 111-136.
[13] Lin, C. H., Wei, A., Yang, H., Pittayachawan, S., Vogel, D., & Wickramasinghe, N. (2015, January). Inquiring Knowledge Management Systems--A Chinese Medicine Perspective. In 2015 48th Hawaii International Conference on System Sciences (pp. 3682-3690). IEEE.
[14] Moscoso-Zea, O., Paredes-Gualtor, J., &Luján-Mora, S. (2018). A holistic view of data warehousing in education. IEEE Access, Vol. 6, No. 1, pp 64659-64673
[15] Nemati, H. R., Steiger, D. M., Iyer, L. S., & Herschel, R. T. (2002). Knowledge warehouse: an architectural integration of knowledge management, decision support, artificial intelligence and data warehousing. Decision Support Systems, Vol. 33, No. 2, pp 143-161.
[16] Palepu, R. B., & Rao, D. K. S. (2012). Meta data quality control architecture in data warehousing. International Journal of Computer Science, Engineering and Information Technology, Vol. 2, No. 4, pp 15-24.
[17] Rahman, N. (2017). An empirical study of data warehouse implementation effectiveness. International Journal of Management Science and Engineering Management, Vol. 12, No. 1, pp 55-63.
[18] Rainardi, V. (2008). Introduction to Data Warehousing. Building a Data Warehouse: With Examples in SQL Server, 1-27. United States, Apress.
[19] Santoso, L. W. (2017). Data warehouse with big data technology for higher education. Procedia Computer Science, Vol. 124, pp 93-99.
[20] Vieira, M., Vieira, J., & Madeira, H. (2008). Towards data security in affordable data warehouse. In 7th European dependable computing conference (pp. 149-157)
[21] Wang, J. (Ed.). (2008). Data warehousing and mining: Concepts, methodologies, tools, and applications: Concepts, methodologies, tools, and applications (Vol. 3). IGI Global

AUTHOR

Abobakr Aljuwaiber is an Associate Professor of Knowledge Management. He holds an MBA and a Ph.D. in knowledge management from the University of East Anglia, UK. He is currently a Vice-Dean of Research and Consulting Institute at Taibah University and used to be a Head of Administrative and Financial Sciences and Technologies Department, and a Vice-Dean at Community College (now Applied College), Taibah University, Saudi Arabia. He has published several peer-reviewed research papers in international journals and conference proceedings. His research interests focus on knowledge management, knowledge sharing, organisational learning, and communities of practice, strategic management.